2007

SPC Applications in Syndromic Surveillance

Fricker, Ronald D. Jr.

Fricker, R.D., Jr., Hegler, B.L., and D.A. Dunfee (2007). SPC Applications in Syndromic Surveillance, 2007 Proceedings of the American Statistical Association [CD-ROM], Alexandria, VA: American Statistical Association, 4035-4044.

http://hdl.handle.net/10945/38742

SPC Applications in Syndromic Surveillance

Ronald D. Fricker, Jr., Matthew C. Knitt, Benjamin L. Hegler, David A. Dunfee, and Cecilia X. Hu
Naval Postgraduate School

Abstract

Syndromic surveillance is the regular collection, analysis, and interpretation of real-time and near-real-time indicators of diseases and other outbreaks by public health organizations. This paper provides an overview of how statistical process control (SPC) methods can be modified for use in syndromic surveillance systems and illustrates one approach for comparing the relative performance of various methods. Specifically, we summarize comparisons we have done between: (1) the C1, C2, and C3 methods implemented in the Early Aberration Reporting System (EARS) versus a CUSUM applied to model-based prediction errors, and (2) two new directionally-sensitive multivariate methods, based on the multivariate CUSUM (MCUSUM) and the multivariate exponentially weighted moving average (MEWMA). Based on our analyses, we found that the CUSUM performed better than the EARS' methods across all of the scenarios we evaluated and, similar to results for the univariate CUSUM and EWMA in classical SPC applications, the directionally-sensitive MCUSUM and MEWMA perform very similarly.

KEY WORDS: Statistical process control, biosurveillance, bioterrorism, early event detection, situational awareness.

1. Introduction

The Centers for Disease Control and Prevention (CDC) as well as many state and local health departments around the United States are developing and fielding *syndromic surveillance* systems (CDC, 2004). Making use of existing health-related data, these health surveillance systems are intended to give early warnings of bioterrorist attacks or other emerging health conditions. See Fricker (2007a), Fricker (2007b), and Fricker and Rolka (2006) for more detailed exposition and discussion.

Biosurveillance systems use a variety of temporal and spatial methods for early event detection, often applying variants of the standard univariate statistical process control (SPC) methods: Shewhart, cumulative sum (CUSUM), and/or exponentially weighted moving average (EWMA) charts. Woodall (2006) provides a comprehensive overview of the application of control charts to health surveillance. Montgomery (2001) is an excellent introduction to these methods in a statistical process control setting. Shmueli and Fienberg (2006) and Shmueli (2006) give a review of these and other methods potentially applicable to early event detection in a biosurveillance setting.

The challenge in applying standard SPC methods is that syndromic surveillance generally violates classical SPC assumptions, particularly the assumption of independent and identically distributed observations. Biosurveillance data is generally autocorrelated and frequently has seasonal periodicities. Further, given the goal of quick detection, methods are usually run on individual observations for which an assumption of normality generally does not apply.

In spite of this, the standard SPC methods are sometimes applied with little modification (see, for example, Fricker, 2007b) and in some cases the methods are modified to attempt to account for the autocorrelation. For example, the CDC's Early Abberation Reporting System (EARS) applies variants of the Shewhart chart (CDC, 2006) which use various moving windows of data to estimate the process mean and standard deviation.

In this paper we present an overview of some of our recent work assessing the performance of various SPC methods that have been modified for the syndromic surveillance problem. Specifically, we compare the performance of a standard univariate CUSUM applied to model-based prediction errors to the EARS methods and then we compare two new directionally-sensitive multivariate methods, based on the multivariate CUSUM (MCUSUM) and the multivariate exponentially weighted moving average (MEWMA). All of the comparisons were conducted using simulated syndromic surveillance data. Because of space limitations, we only present an overview of our work. More detail is available in Fricker, Hegler and Dunfee (2007c) and Fricker, Knitt and Hu (2007d), as well as Dunfee and Hegler (2007) and Hu and Knitt (2007).

This paper is organized as follows. In Section 2 we briefly describe how we simulated the syndromic surveillance data and in Section 3 we introduce the metrics we used to compare between methods. In Section 4 we describe how we conducted the univariate comparisons and provide some example results. In Section 5 we describe the multivariate methods, how we conducted the comparisons, and we provide some example results. Finally, in Section 6 we conclude with an abbreviated summary of our findings.

2. Simulating Syndromic Surveillance Data

In order to compare the various methods, we simulated syndromic surveillance data. This simulated data consisted of a variety of background disease incidence patterns with bioterrorism attacks/natural disease outbreaks (which we will refer to herein simply as "outbreaks") of various sizes and durations overlaid. The simulations of both background disease incidence and outbreaks are purposely idealized depictions designed to capture the main features of syndromic surveillance data.

The simulations were conducted in MatLab 7.1.0.246 using

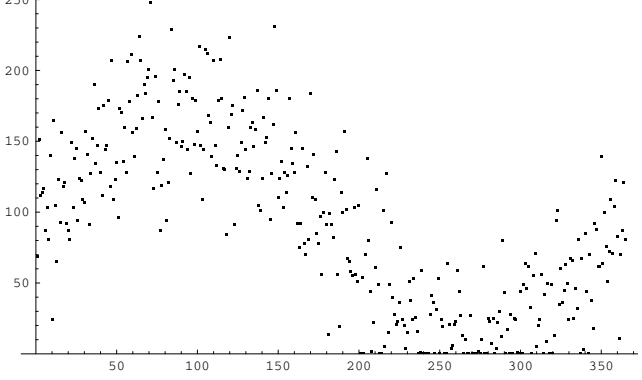

Figure 1: An example of simulated syndromic surveillance data for $c = 90$, $A = 80$, $\mu = 0$ and $\sigma = 30$.

the *randn* function to generate random normal variates and *lognrnd* to generate lognormal random variates.

The background disease incidence data was simulated as the sum of a mean disease incidence, a seasonal sinusoidal cycle, a systematic day-of-the-week effect, and a random fluctuation. Outbreaks, when they occurred, were incorporated as another additive term. That is, a daily observation $Y(t)$ was simulated as

$$Y(t) = \max(0, \lceil c+s(t)+d(t)+Z_c(t)+o(t) \rceil), t = 1, 2, 3, \ldots, \quad (1)$$

where

- c is a constant level of disease incidence;
- s is the seasonal deviation;
- d is the day-of-the-week effect;
- $Z_c(t)$ is the random noise around the systematic component, $c + s(t) + d(t)$;
- $o(t)$ is the mean outbreak level which, when an outbreak is occurring, increases the disease incidence level as described below; and,
- $\lceil x \rceil$ is the ceiling function, which rounds x up to the next largest integer.

The seasonal effect is calculated as $s(t) = A[\sin(2\pi t/365)]$, where A is the maximum deviation from c with $t = 1$ corresponding to October 1st on a 365 day per year calendar. For the random component, we assumed $Z \sim N(\mu, \sigma^2)$ when c is large and $Z \sim LN(\mu, \sigma^2)$ when c is small. The day-of-the-week effect is the systematic deviation from $c + s(t)$, where $d(t) = d(t + 7)$ for all t. It is defined in terms of σ, a parameter of Z.

Parameter values for Equation (1) were pre-specified to define 12 "scenarios" that spanned a range of possible underlying disease incidence patterns. Figure 1 shows an example of the simulated data for $c = 90$, $A = 80$, $\mu = 0$ and $\sigma = 30$.

As described above, outbreaks were incorporated into Equation (1) as an additive term $o(t)$ representing the mean idealized outbreak form that could be parameterized simply, in terms of a peak magnitude M, a duration D, and a random start day τ, where outbreaks started on some day τ, increased linearly up to M over the course of $(D + 1)/2$ days, and then decreased linearly back down to zero.

As we previously mentioned, the characterization of disease incidence in Equation (1) is purposely idealized in order to facilitate comparison of the relative performance of the methods under various scenarios. The idea is to mimic the most salient and important features of syndromic surveillance data in a simulation environment where we can know precisely when outbreaks occur so that we can clearly assess and evaluate performance. That said, it is important to note that the methods do not exploit the idealized features of the data and can be readily adapted to account for those features of real data that are not included in Equation (1). For additional discussion, see Fricker, Hegler and Dunfee (2007c) and Fricker, Knitt and Hu (2007d).

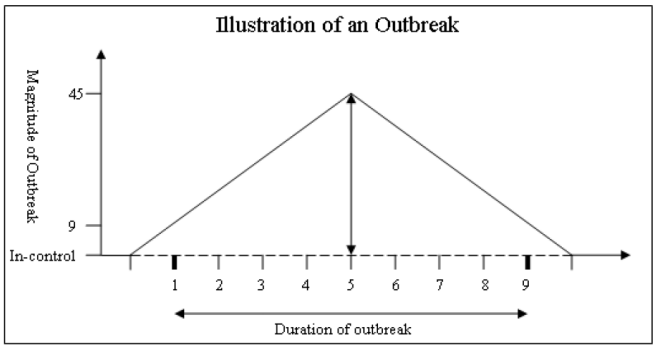

Figure 2: Outbreaks were parameterized in terms of a peak magnitude M and a duration D, where outbreaks started on a random day τ, increased linearly up to M over the course of $(D + 1)/2$ days, and then decreased linearly back down to zero.

3. Comparison Metrics

The syndromic surveillance literature attempts to evaluate performance simultaneously in three dimensions: "sensitivity, specificity, and timeliness." As discussed in Fricker, Hegler and Dunfee (2007) and Fricker, Knitt and Hu (2007), we consider the sensitivity, specificity, and timeliness measures of the syndromic surveillance literature ill-defined.

The metrics we use to compare performance between methods are: (1) the fraction of times a method missed detecting an outbreak and (2) the average time to first outbreak signal (ATFOS). The former is a measure of detection capability while the latter is a conditional measure of the timeliness of detection. The ATFOS is defined as the average time until the first signal among all simulations for which a signal occurred during the outbreak period. Clearly performance in both dimensions must be considered since a desirable method must simultaneously have a short ATFOS and a low fraction of outbreaks missed. A method that is small in one dimension while being

In addition, we set signal thresholds to achieve a specific average time between false signals (ATFS), where the ATFS is determined using a specific background disease incidence pattern during which no outbreaks were allowed. This is similar to the approach used in the SPC literature, where the ATFS is roughly equivalent to the "in-control average run length" and the ATFOS is equivalent to the "out-of-control average run length." The average run length, or ARL, is the average number of observations until a signal. In the SPC literature, it is common practice to compare the performance of methods by first setting thresholds that achieve a specific in-control average run length and then compare out-of-control average run lengths under various conditions. The method that demonstrates lower out-of-control average run lengths across a variety of conditions deemed important is judged to be the better method.

However, our metrics differ from the SPC literature because we also use the fraction missed metric. In the SPC literature, once a process goes out-of-control, it is assumed to stay in that condition until a method signals and the cause is identified and corrected. Thus, once a method goes out of control, any signal is a true signal. This is not the case in syndromic surveillance where outbreaks are transient and after some period of time disappear. In this situation, it is possible for a method to fail to signal during an outbreak, after which a signal is a false signal.

4. Univariate Comparisons: EARS' Methods vs. CUSUM on Adaptive Regression Residuals

In this section we provide an overview of a comparison between the C1, C2, and C3 methods and the cumulative sum (CUSUM) method applied to the prediction errors of a model based on the adaptive regression methodology proposed by Burkom et al. (2007). The need for a model-based approach arises because syndromic surveillance data generally contain uncontrollable trends, periodicities, and day-of-the-week and other effects. See Shmueli (2006) for additional discussion about syndromic surveillance data. See Fricker, Hegler and Dunfee (2007c) and Dunfee and Hegler (2007) for our complete results.

4.1 Methods

4.1.1 EARS' C1, C2, and C3

The C1, C2, and C3 methods were intended to be CUSUM-like methods and, in fact, at least one paper (Zhu et al., 2005) explicitly refers to them as CUSUMs. However, the C1 and C2 are actually Shewhart method variants that use a moving sample average and sample standard deviation to standardize each observation. The C1 method uses the seven days prior to the current observation to calculate the sample average and sample standard deviation. The C2 is similar to the C1 but uses the seven days prior to a two-day lag. The C3 method combines information from C2 statistics as described below.

for example, the number of individuals arriving at a particular hospital emergency room with a specific syndrome on day t. The C1 method calculates the statistic $C_1(t)$ as

$$C_1(t) = \frac{Y(t) - \bar{Y}_1(t)}{S_1(t)} \tag{2}$$

where $\bar{Y}_1(t)$ and $S_1(t)$ are the moving sample mean and standard deviation, respectively:

$$\bar{Y}_1(t) = \frac{1}{7}\sum_{i=t-1}^{t-7} Y(i) \text{ and } S_1^2(t) = \frac{1}{6}\sum_{i=t-1}^{t-7}\left[Y(i) - \bar{Y}_1(i)\right]^2.$$

As implemented in the EARS system, the C1 method signals at time t when the C_1 statistic exceeds a threshold h, which is fixed at three sample standard deviations above the sample mean: $C_1(t) > 3$.

The C2 method is similar to the C1 method, but incorporates a two-day lag in the mean and standard deviation calculations. Specifically, it calculates

$$C_2(t) = \frac{Y(t) - \bar{Y}_3(t)}{S_3(t)} \tag{3}$$

where

$$\bar{Y}_3(t) = \frac{1}{7}\sum_{i=t-3}^{t-9} Y(i) \text{ and } S_3^2(t) = \frac{1}{6}\sum_{i=t-3}^{t-9}\left[Y(i) - \bar{Y}_3(i)\right]^2,$$

and in EARS it signals when $C_2(t) > 3$.

The C3 method uses the C2 statistics from day t and the previous two days, calculating the statistic $C_3(t)$ as

$$C_3(t) = \sum_{i=t}^{t-2} \max\left[0, C_2(i) - 1\right]. \tag{4}$$

In EARS it signals when $C_3(t) > 2$.

In our comparisons between the EARS methods and the CUSUM, we do not use the threshold values given above, but adjust them to achieve an equal ATFS when no outbreak is present. Details on how we calculated the ATFS are described in Fricker, Hegler and Dunfee (2007c).

4.1.2 The CUSUM

The CUSUM method is a well known statistical process control methodology. Formally, the CUSUM is a sequential hypothesis test for a change from a known in-control density f_0 to a known alternative density f_1. The method monitors the statistic $S(t)$, which satisfies the recursion

$$S(t) = \max[0, S(t-1) + L(t)], \tag{5}$$

where the increment $L(t)$ is the log likelihood ratio

$$L(t) = \log\frac{f_1[Y(t)]}{f_0[Y(t)]}.$$

The method is usually started at $S(0) = 0$; it stops and concludes that $Y \sim F_1$ at the first time when $S(t) > h$, for some

when no outbreak is present).

If f_0 and f_1 are normal densities with means μ and $\mu + \delta$, respectively, and unit variances, then Equation (5) reduces to

$$S(t) = \max[0, S(t-1) + Y(t) - \mu - k], \tag{6}$$

where a common choice for k is $k = \delta/2$.

Equation (6) is the form routinely used, even when the underlying data is only approximately normally distributed. It is a one-sided CUSUM, meaning that it will only detect increases in the mean. If it is important to detect both increases and decreases in the mean, a second CUSUM must be used to detect decreases. However, in syndromic surveillance decreases are not relevant since it is only important to quickly detect increases in disease incidence.

In industrial settings, the CUSUM is applied directly to the observations because some control is exhibited over the process such that it is reasonable to assume F_0 is stationary. In syndromic surveillance this is generally not the case as the data often has uncontrollable systematic trends, such as seasonal cycles and day-of-the-week effects. One solution is to model the systematic component of the data, use the model to forecast the next day's observation, and then apply the CUSUM to the forecast errors (Montgomery, 2001).

4.1.3 Applying the CUSUM to Adaptive Regression Residuals

We used the "adaptive regression model with sliding baseline" of Burkom et al. (2007) to model the systematic component of the syndromic surveillance data. The basic idea is as follows. Let $Y(i)$ be an observation, say chief complaint count on day i. Regress the observations for the past n days on time relative to the current period. Then use the model to predict today's observation and apply the CUSUM to the difference between the predicted value and today's observed value. Repeat this process each day, always using the most recent n observations as the sliding baseline in the regression to calculate the forecast error. For $t > n$, and assuming a linear formulation with day-of-the-week effects, the model is

$$\begin{aligned} Y(i) &= \beta_0 + \beta_1 \times (i - t + n + 1) + \beta_2 I_{\text{Mon}} + \beta_3 I_{\text{Tues}} \\ &\quad + \beta_4 I_{\text{Wed}} + \beta_5 I_{\text{Thurs}} + \beta_6 I_{\text{Fri}} + \beta_7 I_{\text{Sat}} + \epsilon \end{aligned} \tag{7}$$

for $i = t-1, \ldots, t-n$. The Is are indicators, where $I = 1$ on the relevant day of the week and $I = 0$ otherwise, and ϵ is the error term which is assumed to follow a symmetric distribution with mean 0 and standard deviation σ_ϵ. Of course, as appropriate, the model can also be adapted to allow for nonlinearities by adding a quadratic term into Equation (7).

Burkom et al. (2007) used an 8-week sliding baseline ($n = 56$). We compared the performance for a variety of n values and between a linear and quadratic form of the model. Fricker, Hegler and Dunfee (2007c) describes how we determined the form for the adaptive regression and the n values.

The model is fit using ordinary least squares, regressing $Y(t-1), \ldots, Y(t-n)$ on $n, \ldots, 1$. Having fit the model,

$$r(t) = Y(t) - \left[\hat{\beta}_0 + \hat{\beta}_1 \times (n+1)\right],$$

where $\hat{\beta}_0$ is the estimated slope and $\hat{\beta}_1$ is the estimated intercept. For any other day of the week the forecast error is

$$r(t) = Y(t) - \left[\hat{\beta}_0 + \hat{\beta}_1 \times (n+1) + \hat{\beta}_j\right]$$

where $\hat{\beta}_2$ is the estimated day-of-the-week effect for Monday, $\hat{\beta}_3$ is for Tuesday, etc.

Standardizing $r(t)$ on σ_ϵ, we have $x(t) = r(t)/\sigma_\epsilon$ and the CUSUM is thus

$$S(t) = \max[0, S(t-1) + x(t) - k], \tag{8}$$

where we assume the expected value of the residuals is zero. (If σ_ϵ is not known, it can be estimated in the usual way from the residuals.) It now remains to determine k.

As shown in the appendix of Fricker, Hegler and Dunfee (2007c), we can estimate the standard deviation of the forecast error for a simple linear adaptive regression as

$$\sigma_{p.e.} = \sigma_\epsilon \sqrt{\frac{(n+2)(n+1)}{n(n-1)}}. \tag{9}$$

Assuming it is important to detect an increase in the mean disease incidence of one standard deviation of the prediction error, then set

$$k = \frac{1}{2}\sqrt{\frac{(n+2)(n+1)}{n(n-1)}},$$

where σ_ϵ (or $\hat{\sigma}_\epsilon$) does not appear in the expression because the CUSUM in Equation (8) uses the standardized residuals. See Fricker, Hegler and Dunfee (2007) for further discussion and the appendix for derivation of Equation (9) as well as the equivalent expression for a multiple regression incorporating day-of-the-week indicator variables.

4.2 Illustrative Univariate Results

Figure 3 in many ways summarizes the results of all the evaluations we conducted. In it, the plots on the left side show the average time to first outbreak signal (ATFOS) versus various outbreak durations (D) in a scenario with $c = 90$, $A = 80$, $\mu = 0$ and $\sigma = 30$, starting with a small outbreak at the top ($M = 9$), a medium outbreak in the middle ($M = 22.5$), and a large outbreak at the bottom ($M = 45$). The plots on the right side show the fraction of times a method missed detecting an outbreak versus outbreak duration. Each plot gives the results for six methods, the C1, C2, and C3, as well as three CUSUMs using various sliding baseline lengths (n values): 7, 15 (the "optimal" for this scenario), and 56 days.

What Figure 3 shows is that the C1, C2, and C3 methods do not perform as well as the CUSUM methods with the larger sliding baselines. Focusing for a moment just on the C1, C2, and C3 methods, we see that the C1 and C2 methods perform

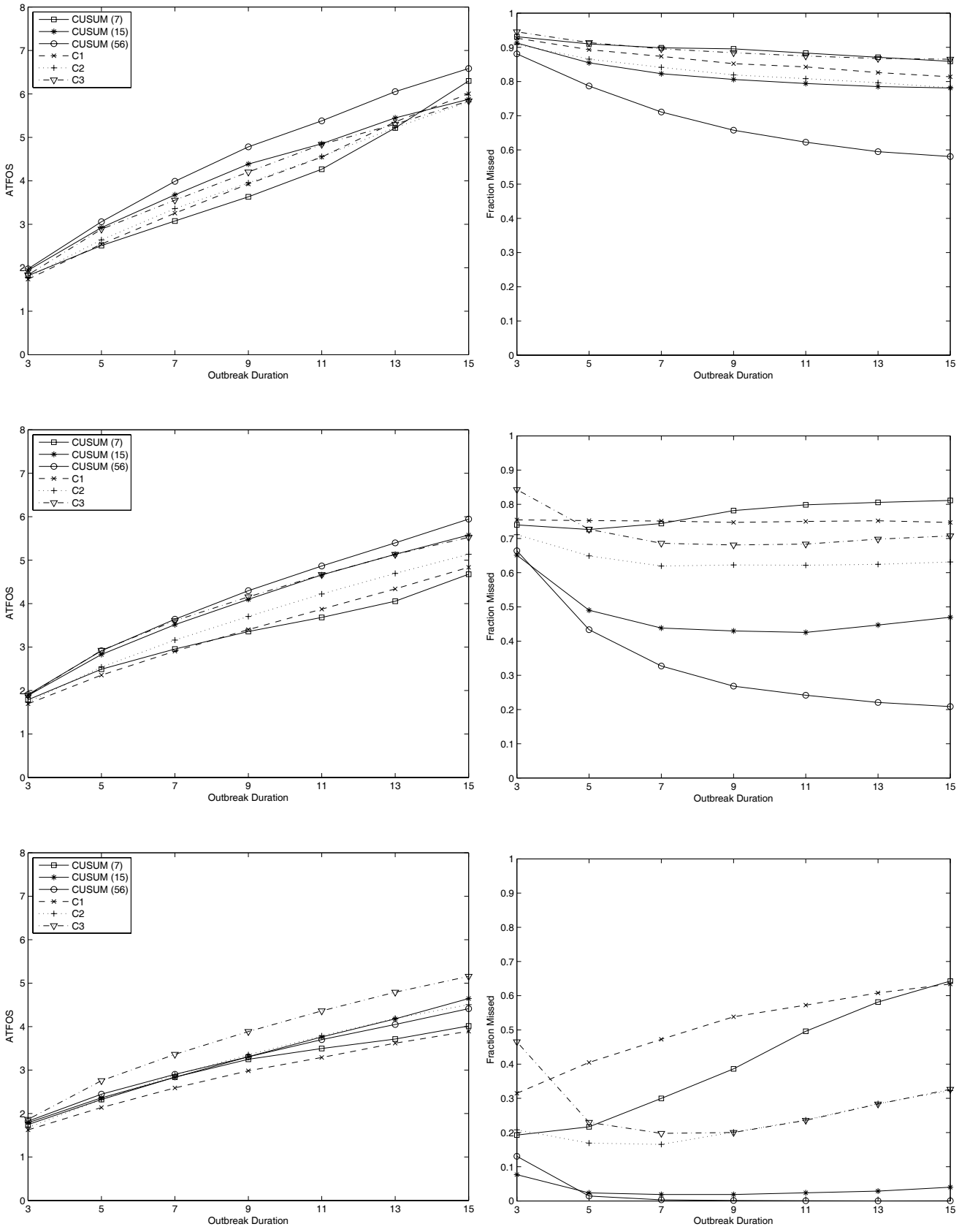

Figure 3: Performance of the methods for a scenario with $c = 90$, $A = 80$, $\mu = 0$ and $\sigma = 30$ for three magnitudes of outbreaks – $M = 9$, $M = 22.5$, and $M = 45$, shown from top to bottom – versus various outbreak durations.

lower ATFOS compared to the C2 but at the expense of having slightly higher fraction missed than the C2. However, when comparing the C1 and C2 to the CUSUMs, we see that they all have similar ATFOS performance but the CUSUMs with longer sliding baselines miss significantly fewer outbreaks. This difference in performance is evident for all the outbreak magnitudes, but is most striking with the larger magnitude outbreaks. For example, in the middle row of plots, the C1 and C2 ATFOS can be up to a day or two shorter than the longer sliding baseline CUSUMs, but they only catch between about 25-35 percent of the outbreaks while the 56-day sliding baseline CUSUM catches nearly 80 percent of the outbreaks of the longest duration. For this scenario, it is clear that the CUSUM with a 56-day sliding baseline is the preferred method.

A note about the ATFOS plots is in order for those used to looking at graphs of average run lengths in the statistical process control literature. Such readers may be surprised that the ATFOS curves increase as outbreak duration increases. In this problem the time to first outbreak signal is constrained to the interval $[1, D]$. That is, the earliest a "true signal" can occur is on the first day of the outbreak and the latest is on the last day of the outbreak (D). Thus, for $D = 3$, ATFOS is constrained to be between 1 and 3 and, as we see in the plot, is about 2 for all the methods. On the other hand, for $D = 15$ the ATFOS can be much larger and, in fact, falls anywhere from about 4 days to about 7 days for the various methods.

Also in Figure 3, we see that the C1 and the CUSUM with a 7-day sliding baseline suffer from being contaminated by the outbreak data in the largest magnitude outbreak scenarios. That is, in the lower right plot we see that the fraction missed by these two methods actually *increases* for longer duration outbreaks (as eventually does the C2 and C3, as well as the CUSUM with a 15-day sliding baseline ever so slightly). If these methods fail to detect the outbreak early on, they begin to incorporate the outbreak data into their calculations (either the moving average for the C1 or the adaptive regression predictions for the CUSUM), making it increasingly more difficult to distinguish the outbreak from the normal background disease incidence. In comparison, the two-day lag in the C2 method seems to be sufficient to eliminate much of this problem for that method (and the C3 which is a function of the C2 statistics).

5. Multivariate Comparisons: MEWMA vs. MCUSUM

In this section we provide an overview of a comparison between two new directionally-sensitive multivariate methods, based on the multivariate CUSUM (MCUSUM) and the multivariate exponentially weighted moving average (MEWMA). While neither of these methods is currently in use in a biosurveillance system, they are among the most promising multivariate methods for this application. The MCUSUM and the MEWMA are also applied to residuals from an adaptive regression that accounts for the systematic effects normally present in syndromic surveillance data. See Fricker, Knitt and Hu (2007c) and Hu and Knitt (2007) for our complete results.

5.1.1 *Directional MCUSUM*

Consider a p-dimensional set of observations at time t, $\mathbf{X}_t = \{X_1, \ldots, X_p\}$. In syndromic surveillance one might think of this as a vector of "chief complaint" counts on day t at p different hospitals in some region. Chief complaints are broad categories – e.g., respiratory, gastrointestinal, unspecified infection, neurological, etcetera – into which patients are grouped *before diagnosis*. Chief complaint is the primary symptom or reason a patient sought care.

Crosier (1988) proposed a MCUSUM that at each time t calculates the statistic

$$\mathbf{S}_t = (\mathbf{S}_{t-1} + \mathbf{X}_t - \boldsymbol{\mu})(1 - k/d_t), \text{ if } d_t > k, \qquad (10)$$

where $\boldsymbol{\mu}$ is the mean of $\mathbf{X}_t$, k is a predetermined statistical distance, and $d_t = [(\mathbf{S}_{t-1}+\mathbf{X}_t-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{S}_{t-1}+\mathbf{X}_t-\boldsymbol{\mu})]^{1/2}$. If $d_t \leq k$ then reset $\mathbf{S}_t = \mathbf{0}$. The method starts with $\mathbf{S}_0 = \mathbf{0}$ and sequentially calculates

$$C_t = (\mathbf{S}_t'\boldsymbol{\Sigma}^{-1}\mathbf{S}_t)^{1/2},$$

where $\boldsymbol{\Sigma}$ is the variance-covariance matrix of $\mathbf{X}$. It concludes that a change has occurred at the first time when $C_t > h$, for some pre-specified threshold h that achieves a desired ATFS.

In terms of choosing $\mathbf{k}$, Crosier (1988) states, "In the univariate [CUSUM] case, the quantity $S_{t-1}+(X_t-\mu)$ is shrunk towards 0 by k standard deviations. If this is to hold for the multivariate case, $\mathbf{k}$ must satisfy $\mathbf{k}'\boldsymbol{\Sigma}^{-1}\mathbf{k} = k^2$ – that is, $\mathbf{k}$ must be of length k, where the length is defined by using the covariance matrix $\boldsymbol{\Sigma}$."

The literature contains a number of MCUSUM methods. In fact, the Crosier method described above is one of a number of other multivariate CUSUM-like methods he proposed, but Crosier generally preferred the above method after extensive simulation comparisons. Pignatiello and Runger (1990) proposed other multivariate CUSUM-like methods but found that they performed similar to Crosier's. Healy (1987) derived a sequential likelihood ratio test to detect a shift in a mean vector of a multivariate normal distribution. However, while Healy's method is more effective when the change is to the precise mean vector to be detected, it is less effective than Crosier's for detecting other types of shifts, including mean shifts that were close to but not precisely the specified mean vector.

For the syndromic surveillance problem, an advantage of Crosier's MCUSUM formulation is that it is easy to modify to only look for positive increases. As described in Fricker (2007b), the motivation for this modification is the univariate CUSUM where directionality is achieved because the CUSUM statistic is bounded below by zero. In the modified MCUSUM directionality is similarly achieved by bounding each component of the cumulative sum vector by zero. In particular, for detecting positive increases relevant to the syndromic surveillance problem, when $d_t > k$ limit $\mathbf{S}_t$ to be nonnegative in each dimension by replacing Equation (10) with $\mathbf{S}_t = (S_{t,1}, \ldots, S_{t,p})$ where

$$S_{t,j} = \max[0, (S_{t-1,j} + X_{t,j} - \mu_j)(1 - k/d_t)],$$

for $j = 1, 2, \ldots, p$.

Lowry et al. (1992) introduced the MEWMA as a generalization of the univariate EWMA of Roberts (1959). As with the MCUSUM, denote the mean for $\mathbf{X}_t$ as $\boldsymbol{\mu}$ and let $\boldsymbol{\Sigma}$ be the covariance matrix. In the spirit of the reflected EWMA of Crowder and Hamilton (1992), the directionally sensitive MEWMA proposed by Joner et al. (2007) calculates

$$\mathbf{Z}_t = \begin{cases} \max[\mathbf{0}, \lambda(\mathbf{X}_t - \boldsymbol{\mu}) + (1-\lambda)\mathbf{Z}_{t-1}], & \text{for } t > 0 \\ \mathbf{0}, & \text{for } t = 0 \end{cases},$$

where the maximum function is applied componentwise. $\mathbf{Z}_t$ is a weighted average of the current observation standardized around $\mathbf{0}$ and the previous $\mathbf{Z}$ statistic. The parameter $0 < \lambda \leq 1$ is the *smoothing parameter* which controls the weight assigned to the new observation vector. The covariance matrix for $\mathbf{Z}_t$ is

$$\boldsymbol{\Sigma}_{\mathbf{Z}_t} = \frac{\lambda\left[1-(1-\lambda)^{2t}\right]}{2-\lambda}\boldsymbol{\Sigma}.$$

Taking the limit as $t \to \infty$, we have

$$\boldsymbol{\Sigma}_{\mathbf{Z}_\infty} = \frac{\lambda}{2-\lambda}\boldsymbol{\Sigma}.$$

$\boldsymbol{\Sigma}_{\mathbf{Z}_\infty}$ is then used to calculate the MEWMA test statistic E_t where

$$E_t = \mathbf{Z}_t' \Sigma_{\mathbf{Z}_\infty}^{-1} \mathbf{Z}_t.$$

The MEWMA signals an alarm whenever E_t exceeds a predetermined threshold h which is set to achieve a desired ATFS. If E_t does not exceed h, then the MEWMA iterates through the next time step with a new observation vector, recalculating the test statistic, and continuing until such time as the $E_t > h$.

5.1.3 Setting Parameters

In order to compare the MEWMA and MCUSUM under a variety of syndromic surveillance scenarios, we first fixed λ for the MEWMA and then searched for the value of k in the MCUSUM that matched its performance to the MEWMA's. Montgomery (2001) recommends setting $0.05 \leq \lambda \leq 0.25$ for the univariate EWMA and, given the emphasis on timeliness in this application and based on our experience, we thus chose to set $\lambda = 0.2$. Having fixed λ, we conducted simulation comparisons over various values of k to find that value for which the MCUSUM performed as closely as possible to the MEWMA. We found that $k = 0.74$ gave the closest performance to the MEWMA with $\lambda = 0.2$. Please see Fricker, Knitt and Hu (2007) or Hu and Knitt (2007) for additional details.

5.2 Illustrative Multivariate Results

Figures 4 and 5 summarize our main finding: the MEWMA and MCUSUM performed virtually identically, both in terms of ATFOS and percent missed, across all the scenario and outbreak combinations we evaluated. Though the lines deviate slightly in Figures 4 and 5, the differences are not statistically significant. See Hu and Knitt (2007) for details.

MCUSUM and MEWMA performance for the scenario with $c = 90$, $A = 90$, $\mu = 0$, and $\sigma = 10$ across all the types of outbreaks, from small to large magnitudes, and for all the durations. This result was also true for the other scenarios we evaluated. For example, Figure 5 shows the results for three different scenarios for an outbreak of medium magnitude. See Hu and Knitt (2007) for plots for all of the scenarios and types of outbreaks.

Figure 4 demonstrates how the methods perform for the various types of outbreaks. For example, the ATFOS plots show that outbreaks of small magnitude and of three days duration will only be detected about 30 percent of the time and, when detected, it will take about two days on average for either the MCUSUM or MEWMA to signal. As the outbreak magnitude increases, the methods detect virtually all of the outbreaks and the ATFOS decreases to about one day for the largest magnitude outbreak. In comparison, for durations of 15 days, the methods detect almost 70 percent of the small magnitude outbreaks and again virtually all of the larger outbreaks. For the small magnitude outbreaks the average time to signal is about six days, for the medium magnitude it is just under five days, and for the large magnitude outbreak it is about 2-1/2 days.

Figure 5 demonstrates that the adaptive regression with sliding baseline methodology does very well at removing the systematic component, at least for our synthetic syndromic surveillance data. In this case, the systematic component is the seasonal sinusoid where, at the top the sinusoid is large ($A = 90$), in the middle it is medium sized ($A = 20$), and at the bottom it is non-existent ($A = 0$). In terms of ATFOS, there is no visible difference between the three plots in Figure 5. In terms of percent of outbreaks missed, there is a slight degradation in the number of outbreaks caught as the amplitude increases. However, these plots demonstrate that, overall, the adaptive regression is quite effective at accounting for the systematic trends in the data.

6. Discussion

Based on our comparisons of the EARS methods to the CUSUM methods applied to the residuals of adaptive regressions, we find that the CUSUM methods perform better. In particular, the EARS methods frequently failed to catch a majority of the outbreaks across a wide variety of background disease incident patterns (large and small daily counts; large, medium, small, and no seasonal cycles; large and small random daily fluctuations; with and without day-of-the-week effects) and a wide variety of outbreak magnitudes and durations. In fact, the EARS methods generally caught less than 30 percent of the outbreaks except in the largest outbreak cases. In contrast, the CUSUM methods, particularly with the 8-week sliding baseline, performed much better.

Of course, the EARS methods were originally designed for a drop-in surveillance system with little or no baseline data available. In these situations the use of an 8-week sliding baseline may be impossible, at least upon initiation of the drop-in system. However, our simulations showed that a CUSUM with a 7-day sliding baseline performed about the same as the

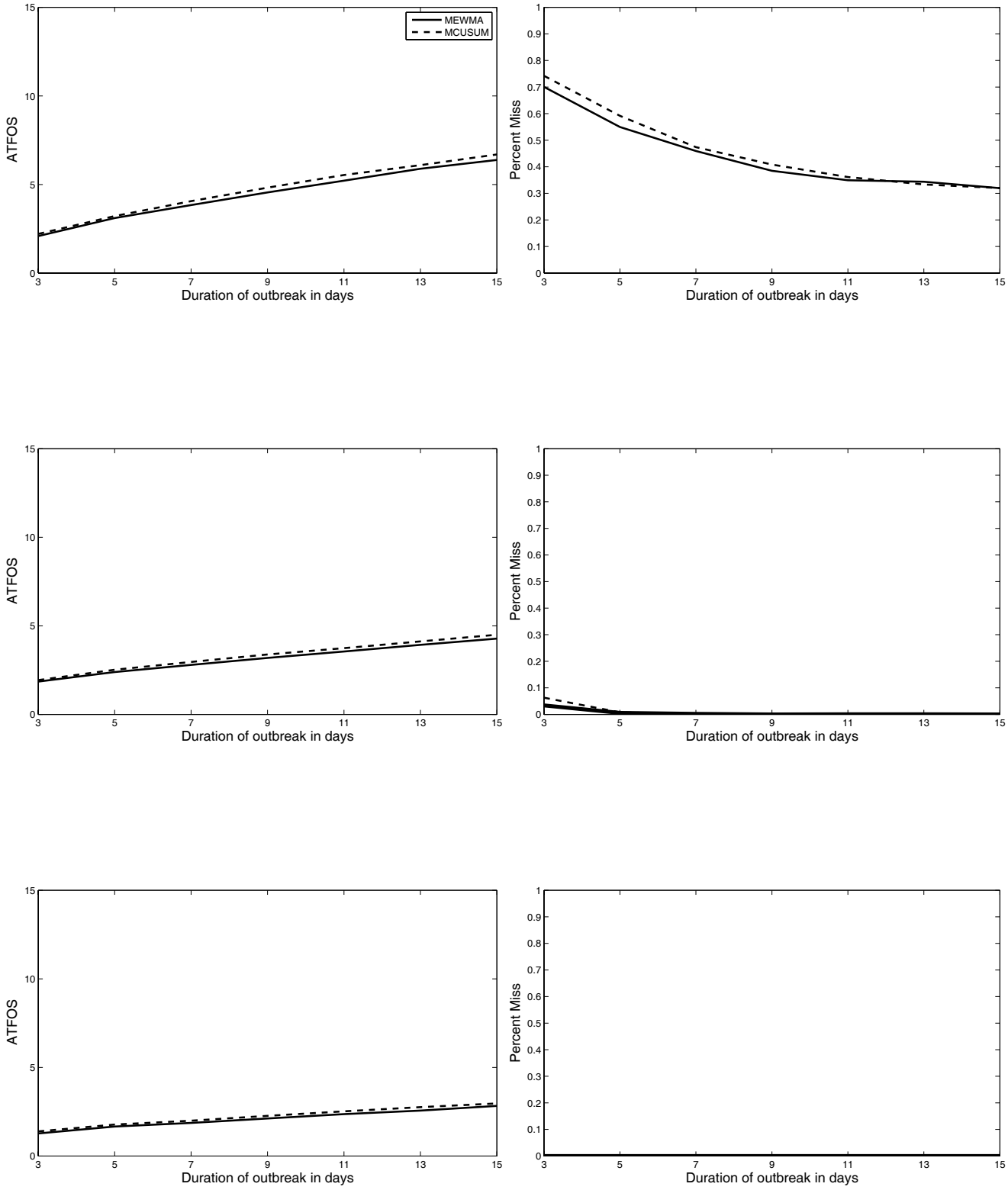

Figure 4: Performance of the MCUSUM and MEWMA under a scenario with $c = 90$, $A = 90$, $\mu = 0$, and $\sigma = 10$ for three magnitudes of outbreaks – $M = 9$, $M = 22.5$, and $M = 45$, shown from top to bottom – versus various outbreak durations.

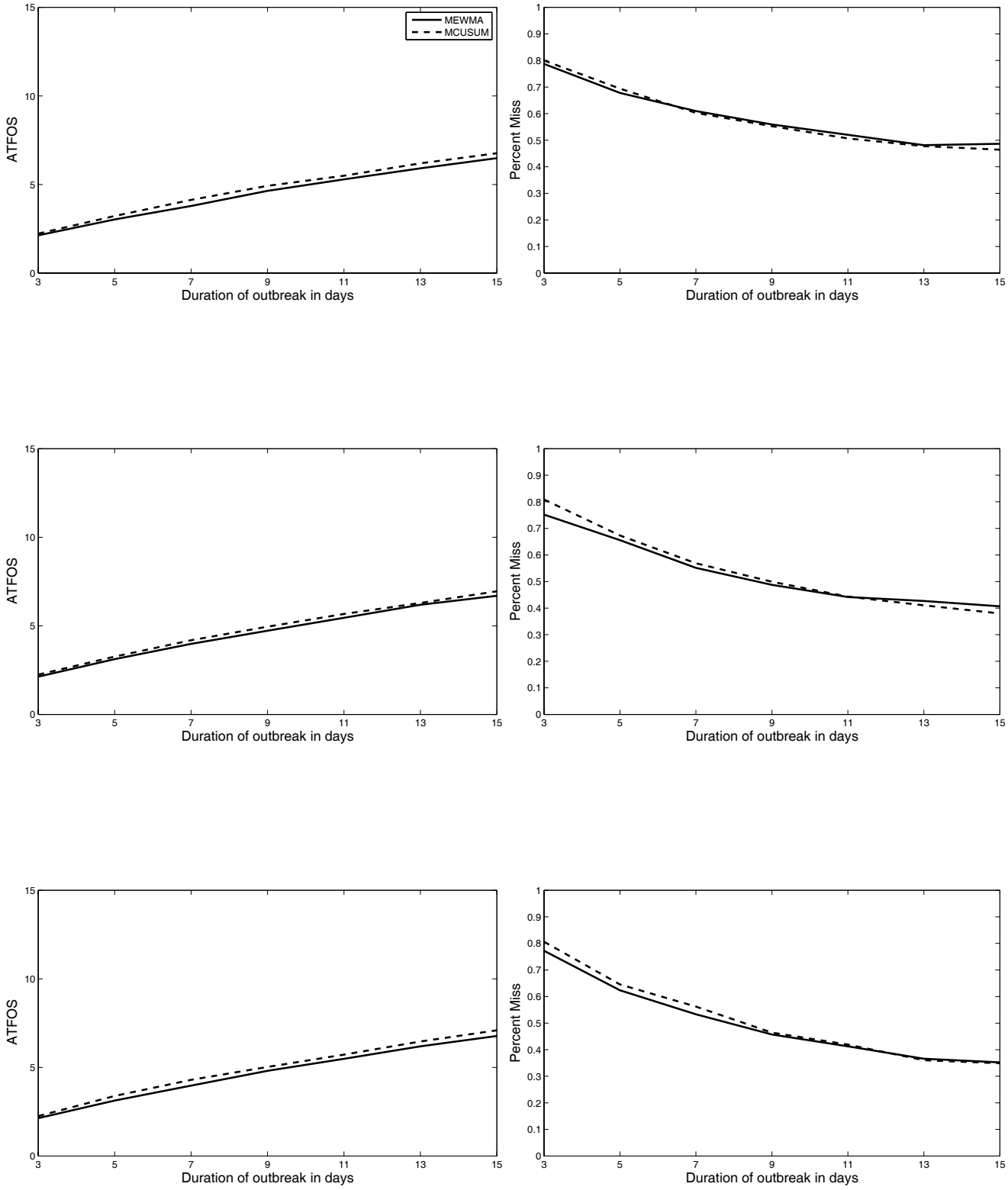

Figure 5: Performance of the MEWMA and MCUSUM for $c = 90$, $\mu = 0$, $\sigma = 30$, and outbreak magnitude $M = 22.5$ for three magnitudes of amplitude – $A = 90$, $A = 20$, and $A = 0$, shown from top to bottom – versus various outbreak durations.

creased the performance of the CUSUM quickly improved. This suggests a strategy for drop-in surveillance systems of starting with a CUSUM with a 7-day sliding baseline and, as time progresses and more data accumulates, allowing the baseline to increase until such time as enough data is accumulated so that baseline can be allowed to slide.

In terms of the multivariate comparisons, when we began this research we fully expected to identify scenarios in which the MCUSUM performed better than the MEWMA and vice versa. That the two methods performed practically identically is an unexpected surprise. It is a surprise because, while it is well-known that with the appropriate choice of parameters the univariate EWMA and CUSUM can be made to perform similarly in standard SPC applications, the directional MEWMA and MCUSUM described herein are neither the exact multivariate analogues of their univariate counterparts nor is the syndromic surveillance problem the same as the standard SPC application.

Because there is seemingly no performance advantage in using one method over the other, this result leads us to prefer the MEWMA for procedural reasons. Specifically, it is relatively easy to develop an intuitive appreciation for how to choose λ and much more difficult to understand how to appropriately choose $\mathbf{k}$. That is, unlike the k in the univariate CUSUM which has a clear interpretation, namely it is one-half of the smallest mean shift that is to be detected quickly, the k in Crosier's MCUSUM is a parameter in a multiplicative "shrinkage factor" for which there is no literature or research to guide one in the trade-offs that must result from various choices of k.

Finally, in all of our analyses, we found the adaptive regression methodology to be effective at removing the systematic effects from the background disease incidence. See Fricker, Hegler and Dunfee (2007c) and Fricker, Knitt and Hu (2007d), as well as Dunfee and Hegler (2007) and Hu and Knitt (2007) for more detail. The longer sliding baseline, along with the linear form of the model, is also effective at ensuring the adaptive regression does not get contaminated in longer duration outbreaks if it uses some of that data in the regression model.

Future research should assess whether these conclusions continue to hold on actual syndromic surveillance data.

Acknowledgements

R. Fricker's work on this effort was partially supported by Office of Naval Research grant N0001407WR20172.

REFERENCES

Burkom, H.S., S.P. Murphy, and G. Shmueli (2007). "Automated Time Series Forecasting for Biosurveillance," *Statistics in Medicine*, **26**, pp. 4202–4218.

The Centers for Disease Control and Prevention (2006). Early Aberration Reporting System Website, www.bt.cdc.gov/surveillance/ears/, accessed on November 22, 2006.

Centers for Disease Control and Surveillance, "Syndromic Surveillance: Reports from a National Conference, 2003," (2004). *Morbidity and Morality Weekly Report*, **53** (Supplement), September.

ity Control Schemes, *Technometrics*, **30**, pp. 291–303.

Crowder, S.V. and Hamilton, M.D. (1992). An EWMA for Monitoring a Process Standard Deviation, *Journal of Quality Technology*, **24**, pp. 12–21.

Dunfee, D.A., and B.L. Hegler (2007). *Biological Terrorism Preparedness: Evaluating the Performance of the Early Aberration Reporting System (EARS) Syndromic Surveillance Algorithms*, Master's Thesis, Naval Postgraduate School, Monterey, CA.

Fricker, R.D., Jr. (2007a). Syndromic Surveillance, *Encyclopedia of Quantitative Risk Assessment* (to appear).

Fricker, R.D., Jr. (2007b). Directionally Sensitive Multivariate Statistical Process Control Methods with Application to Syndromic Surveillance, *Advances in Disease Surveillance*, www.isdsjournal.org, **3**:1. Available on-line at www.isdsjournal.org.

Fricker, R.D., Jr., Hegler, B.L., and D.A. Dunfee (2007c). Comparing Biosurveillance Detection Methods: EARS' Versus a CUSUM-based Methodology, in submission to *Statistics in Medicine*.

Fricker, R.D., Jr., Knitt, M.C., and C.X. Hu (2007d). Directionally Sensitive MCUSUM and MEWMA Procedures with Application to Biosurveillance, in submission to *Quality Engineering*.

Fricker, R.D., Jr., and H. Rolka (2006). Protecting Against Biological Terrorism: Statistical Issues in Electronic Biosurveillance, *Chance*, **91**, pp. 4–13.

Healy, J.D. (1987). A Note on Multivariate CUSUM Procedures, *Technometrics*, **29**, pp. 409–412.

Hu, C.X. and M.C. Knitt (2007). *A Comparative Analysis of Multivariate Statistical Detection Methods Applied to Syndromic Surveillance*, Master's Thesis, Naval Postgraduate School, Monterey, CA.

Hutwagner, L., Thompson, W., Seeman, G.M., and Treadwell, T. (2003). The Bioterrorism Preparedness and Response Early Aberration Reporting System (EARS), *Journal of Urban Health: Bulletin of the New York Academy of Medicine*, **80**, (No. 2, Supplement 1), pp. 89i-96i.

Joner, M.D., Jr., Woodall, W.H., Reynolds, M.R., Jr., and R.D. Fricker, Jr. (2007). A One-Sided MEWMA Chart for Health Surveillance, in submission to *Quality and Reliability Engineering International*.

Lowry, C.A., Woodall, W.H., Champ, C.W., and S.E. Rigdon (1992). A Multivariate Exponentially Weighted Moving Average Control Chart, *Technometrics*, **34**, 1, pp. 46–53.

Montgomery, D.C. (2001). *Introduction to Statistical Quality Control*, 4th edition, John Wiley & Sons, New York.

Pignatiello, J.J., Jr., and G.C. Runger (1990). Comparisons of Multivariate CUSUM Charts, *Journal of Quality Technology*, **3**, pp. 173–186.

Roberts, S.W. (1959). Control Chart Tests Based on Geometric Moving Averages, *Technometrics*, **1**, pp. 239–250.

Shmueli, G. (2006). Statistical Challenges in Modern Biosurveillance, under revision for *Technometrics*, draft dated September 18, 2006.

Shmueli, G., and S.E. Fienberg (2006). Current and Potential Statistical Methods for Monitoring Multiple Data Streams for Biosurveillance, *Statistical Methods in Counterterrorism: Game Theory, Modeling, Syndromic Surveillance, and Biometric Authentication*, A. Wilson, G. Wilson, and D.H. Olwell, eds., Springer, New York, NY.

Woodall, W.H. (2006). The Use of Control Charts in Health-Care and Public-Health Surveillance, *Journal of Quality Technology*, **38**, pp. 1-16. Available at www.asq.org/pub/jqt.

Zhu, Y., Wang, W., Atrubin, D., and Y. Wu (2005). Initial Evaluation of the Early Aberration Reporting System — Florida, *Morbidity and Mortality Weekly Report*, **54** (Supplement), Centers for Disease Control and Prevention, pp. 123–130.